Machine Learning in Biosciences

Instructor: Peng Qiu

Biomedical Engineering
Georgia Tech and Emory

Logistics

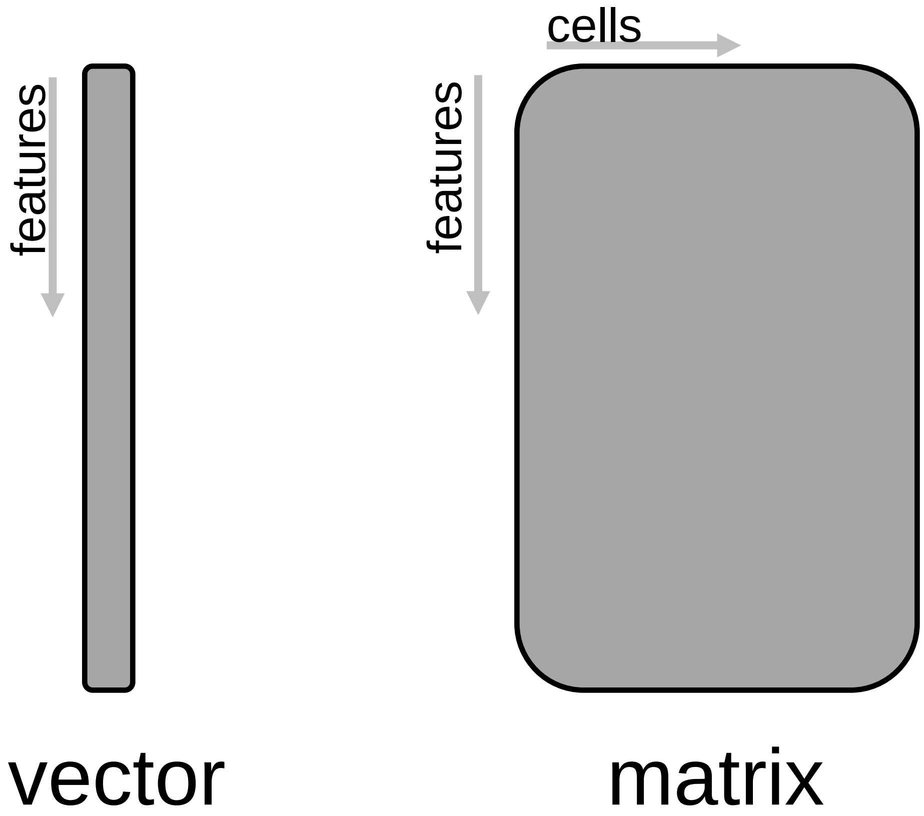
Class attendance tracked from week 02

Project done in teams of 1~3 students.

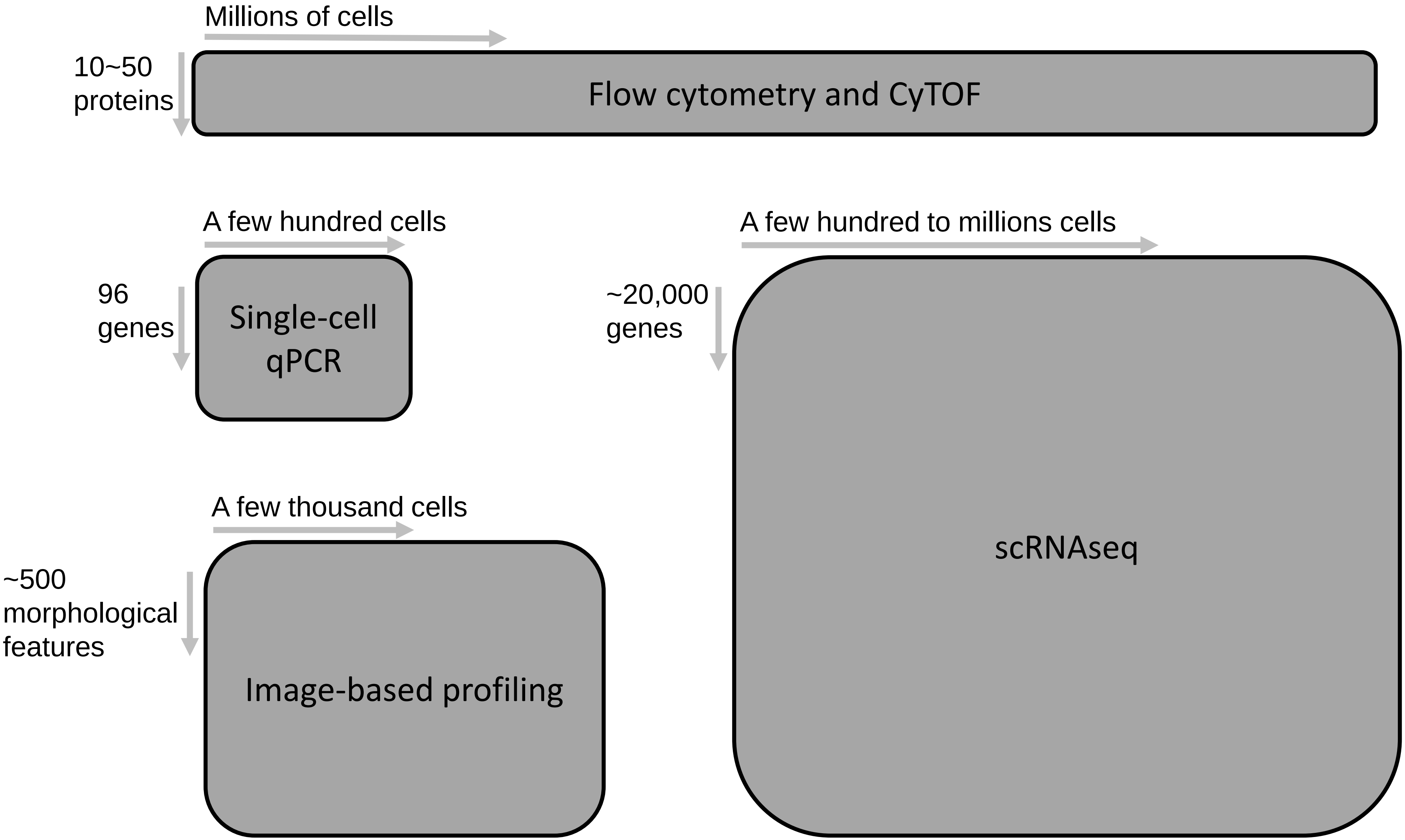
Project does not have to be biology related, but a biology focus is preferred.

Public data resources.

Bulk-tissue vs. Single-cell



A few single-cell technologies



Millions of cells

10~50 proteins

Flow cytometry and CyTOF

A few hundred cells

96 genes

Single-cell qPCR

A few thousand cells

~500 morphological features

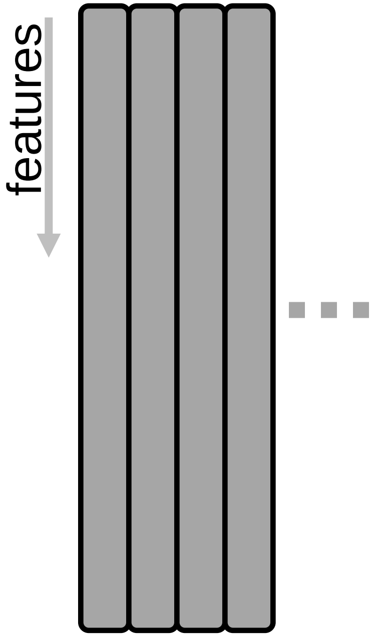
Image-based profiling

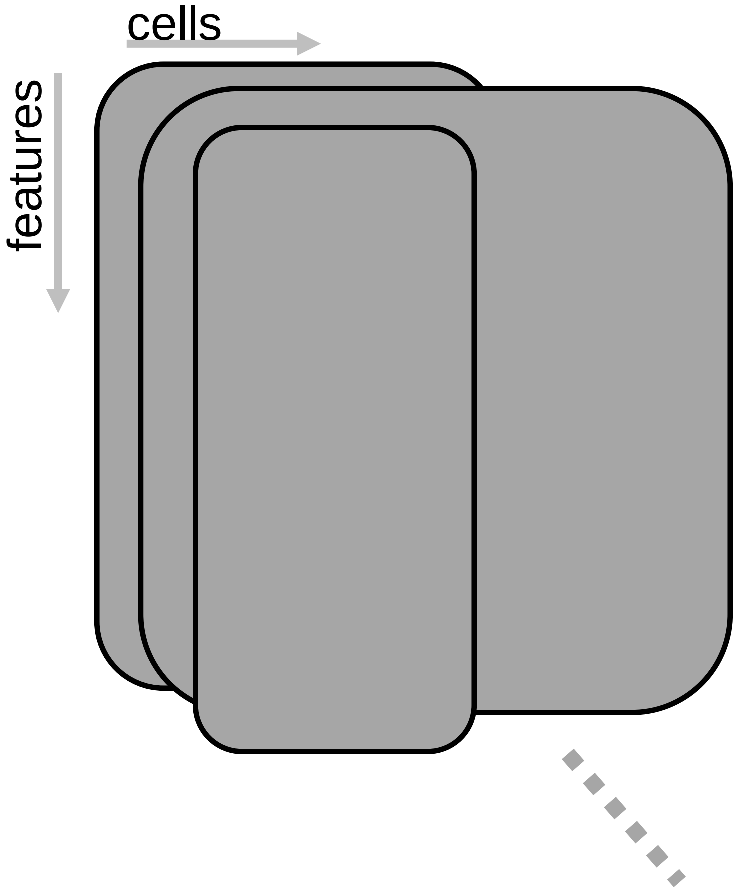
A few hundred to millions cells

~20,000 genes

scRNAseq

Bulk-tissue vs. Single-cell





Early-term Project

Classification of AML

The goal of this project is to predict patient’s AML or normal status from patient blood samples profiled by flow cytometry.

Flow cytometry

- Flow: what is it?
  - Evaluate cells in single-cell suspension
  - Sample is prepared in liquid form
  - Cells flow in a thin stream (usually saline)
  - Cells pass a detector one by one.

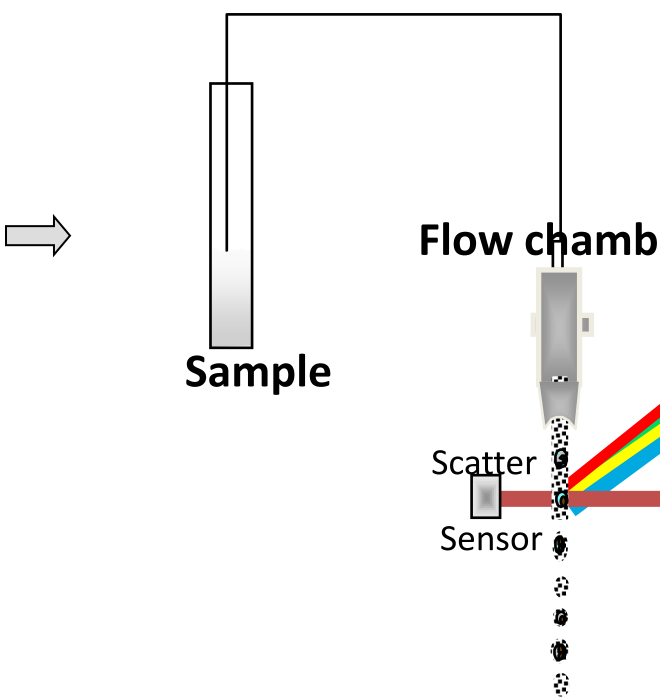
- Cytometry: what does the detector measure?
  - How much of something exists inside or on the surface of a cell?
  - Surface protein markers: CD19/20, CD3/4/8, …
  - Proteins markers inside: pStat3/5, pAKT, …
  - Size, granularity, DNA content, viability …

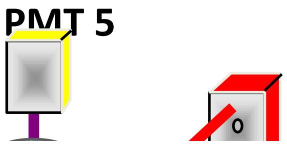
Flow cytometry

Decide what to measure:

	-	'FSC‘
	-	'SSC‘
'Kappa‘	-	'FITC‘
'Lambda‘	-	'PE‘
'CD45‘	-	‘ECD‘
'CD19‘	-	‘PC5‘
'CD20'	-	‘PC7‘

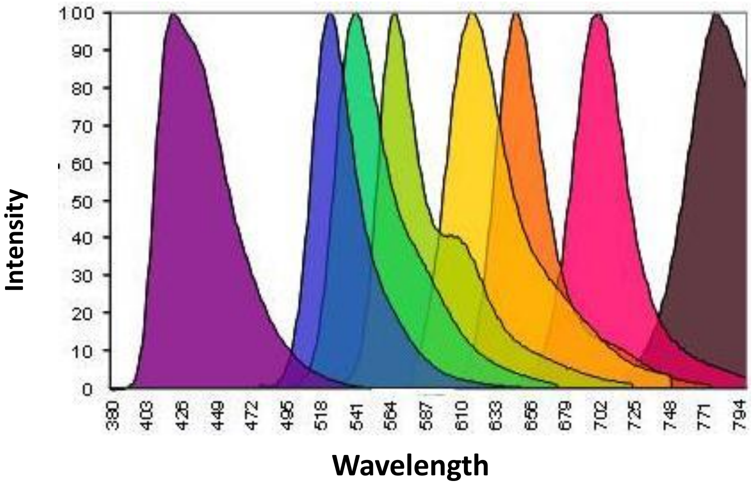
(<= 12)





FSC	SSC	Kappa	Lambda	CD45	CD19	CD20
830	597	407	406	559	43	150
391	386	71	85	624	0	0
1023	868	614	640	409	481	494
571	618	438	425	557	32	59
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Limitation

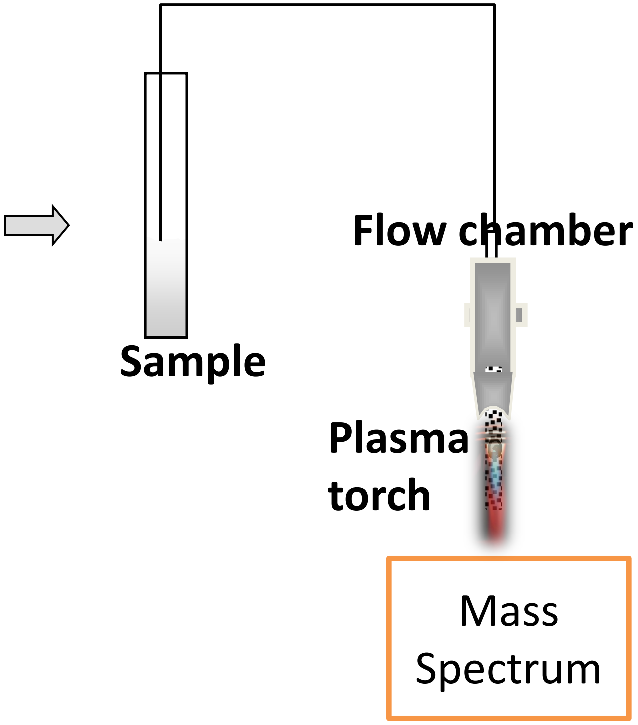

100
90
80
70
60
50
40
30
20
10
0
Intensity
380
403
426
449
472
495
518
541
564
587
610
633
656
679
702
725
748
771
794
Wavelength

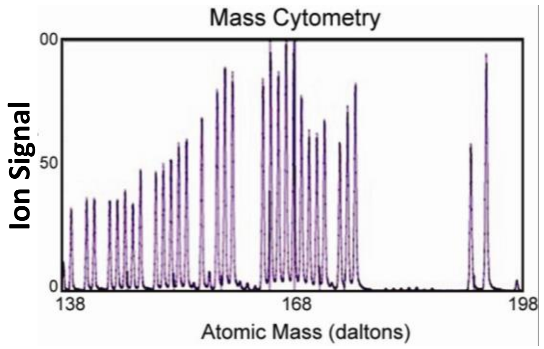
Mass cytometry (CyTOF)

Decide what to measure:

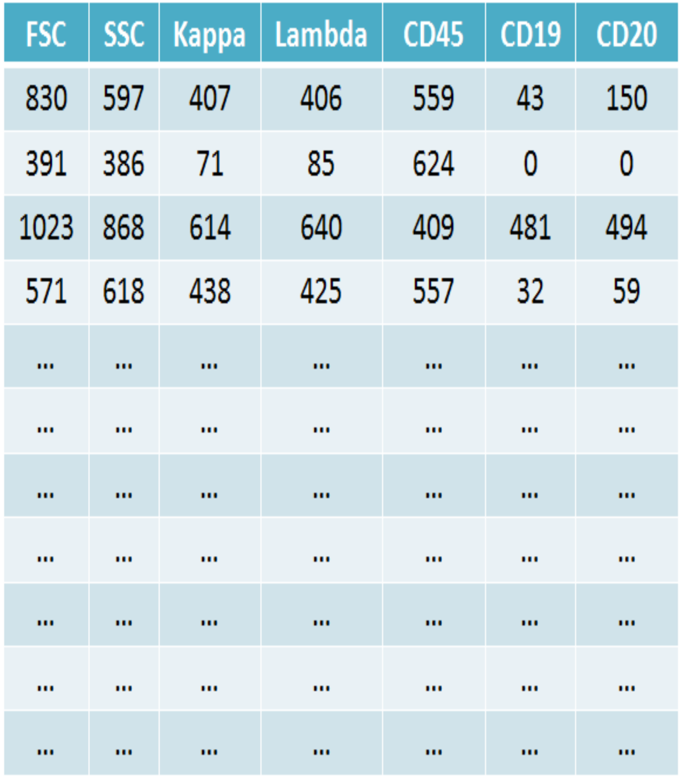
	-	DNA
	-	Viability
'Kappa‘	-	La(138.906)
'Lambda‘	-	Nd(144.912)
'CD45‘	-	Nd(145.913)
'CD19‘	-	Eu(150.919)
'CD20‘	-	Gd(155.922)
…	-	…
…	-	…
…	-	…

(<= 100)





Flow cytometry data for one sample



FSC	SSC	Kappa	Lambda	CD45	CD19	CD20
830	597	407	406	559	43	150
391	386	71	85	624	0	0
1023	868	614	640	409	481	494
571	618	438	425	557	32	59
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Flow cytometry data for one sample

Many many cells

Features

CD20	150	0	494	59	⋮	⋮	⋮	⋮	⋮	⋮	⋮
CD19	43	0	481	32	⋮	⋮	⋮	⋮	⋮	⋮	⋮
CD45	559	624	409	557	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Lambda	406	85	640	425	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Kappa	407	71	614	438	⋮	⋮	⋮	⋮	⋮	⋮	⋮
SSC	597	386	868	618	⋮	⋮	⋮	⋮	⋮	⋮	⋮
FSC	830	391	1023	571	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Biology questions

- Relationship among cells
  - Identify cell types
  - Infer how different cell types are related

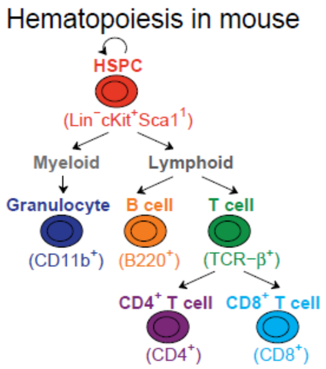
- Relationship among markers
  - Identify signaling network

- Relationship between cells and overall phenotype
  - cellular composition of a tumor vs. survival, (drug response, …)

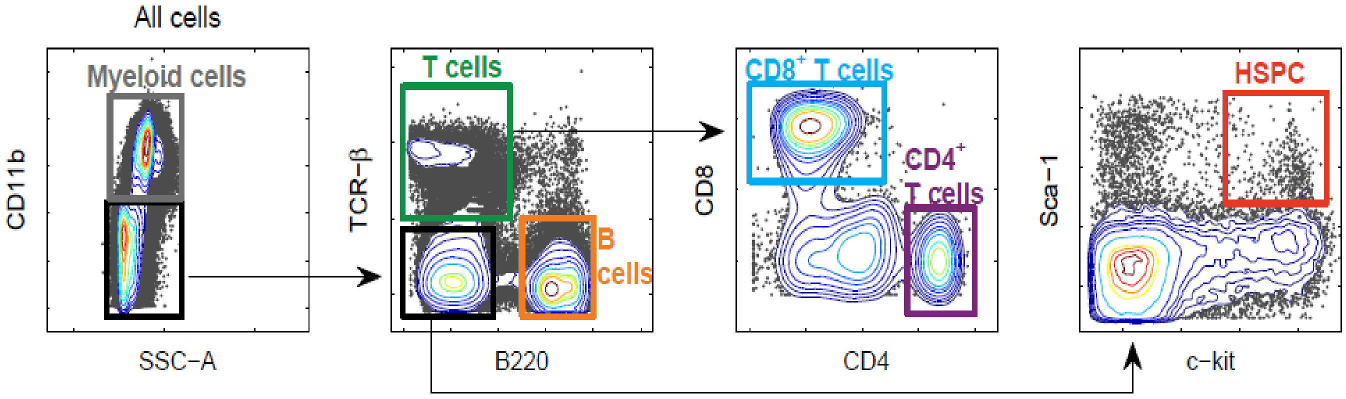
- Relationship between markers and overall phenotype
  - whether a signaling pathway is cell type specific, or disease specific?

Conventional analysis of flow cytometry data

- Example data
  - Flow cytometry
  - Mouse bone marrow
  - Parameters: c-kit, Sca-1, CD11b, B220, TCR-b, CD4, CD8



- Traditional analysis: Gating



(Kenny, Nolan lab)

Flow cytometry data for one sample

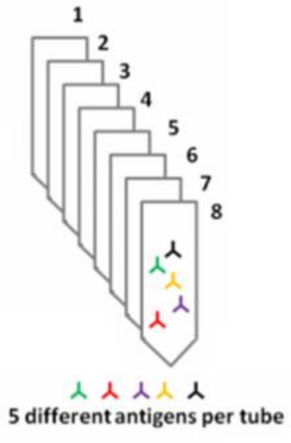
FSC	SSC	Kappa	Lambda	CD45	CD19	CD20
830	597	407	406	559	43	150
391	386	71	85	624	0	0
1023	868	614	640	409	481	494
571	618	438	425	557	32	59
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

FSC	SSC	Kappa	Lambda	CD45	CD19	CD20
830	597	407	406	559	43	150
391	386	71	85	624	0	0
1023	868	614	640	409	481	494
571	618	438	425	557	32	59
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Classification of AML Project

- 359 subjects
  - 316 normal subjects
  - 43 AML samples

- For each subject, one blood sample is taken, and split into 8 tubes.

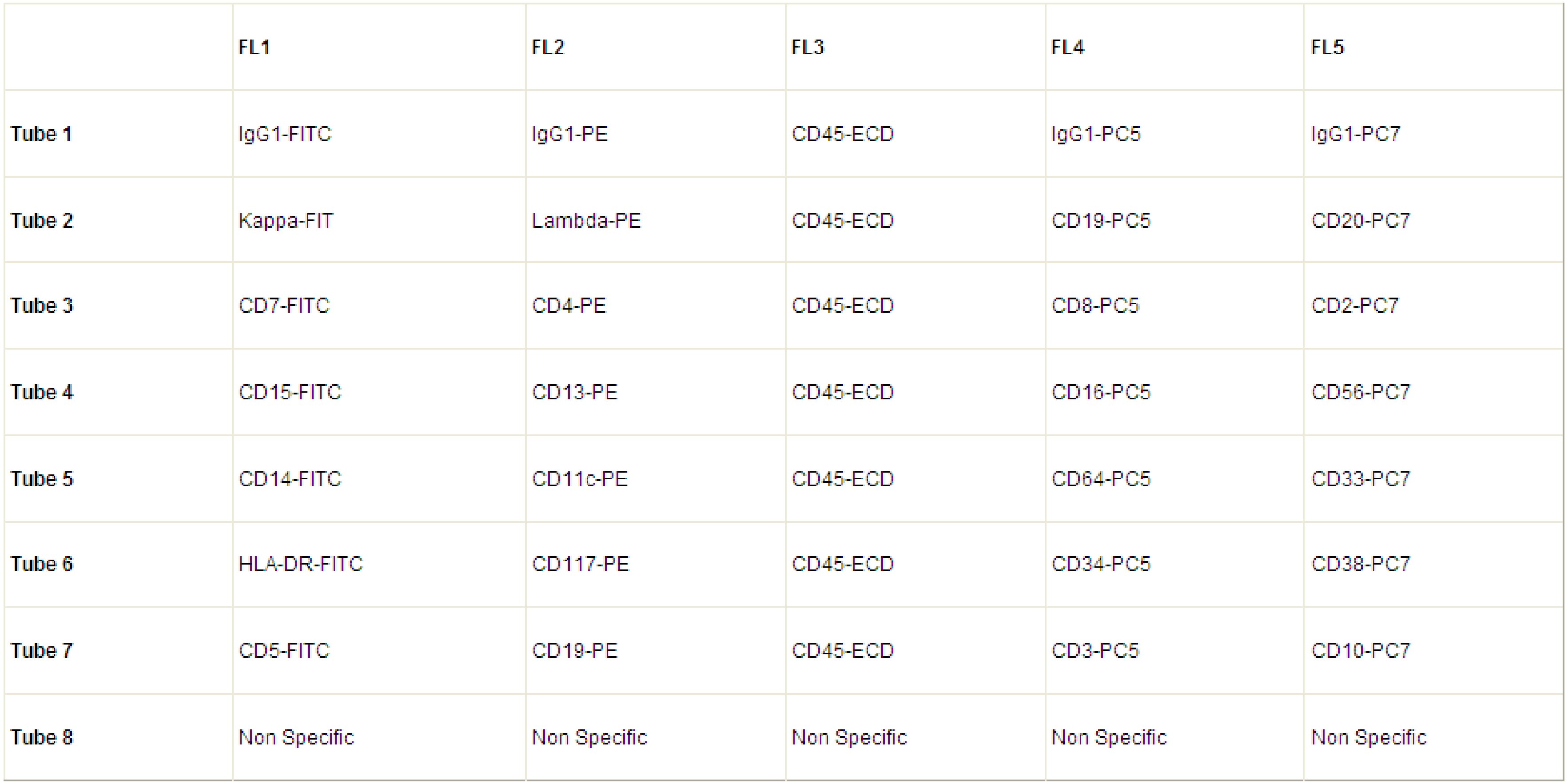


- For each tube, 7 channels are measured
  - FSC, SSC
  - 5 protein markers

- Normal/AML class labels of 179 samples are given

- Task: predict the class labels of the remaining 180 samples

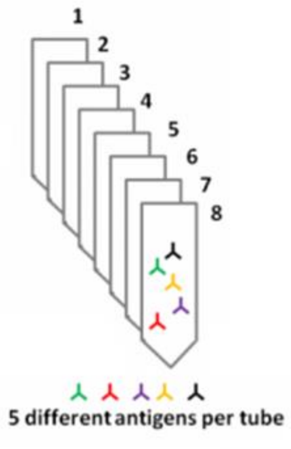
Classification of AML Project



	FL1	FL2	FL3	FL4	FL5
Tube 1	IgG1-FITC	IgG1-PE	CD45-ECD	IgG1-PC5	IgG1-PC7
Tube 2	Kappa-FIT	Lambda-PE	CD45-ECD	CD19-PC5	CD20-PC7
Tube 3	CD7-FITC	CD4-PE	CD45-ECD	CD8-PC5	CD2-PC7
Tube 4	CD15-FITC	CD13-PE	CD45-ECD	CD16-PC5	CD56-PC7
Tube 5	CD14-FITC	CD11c-PE	CD45-ECD	CD64-PC5	CD33-PC7
Tube 6	HLA-DR-FITC	CD117-PE	CD45-ECD	CD34-PC5	CD38-PC7
Tube 7	CD5-FITC	CD19-PE	CD45-ECD	CD3-PC5	CD10-PC7
Tube 8	Non Specific	Non Specific	Non Specific	Non Specific	Non Specific

Classification of AML Project

- 359 subjects
  - 316 normal subjects
  - 43 AML samples



- For each subject, one blood sample is taken, and split into 8 tubes.

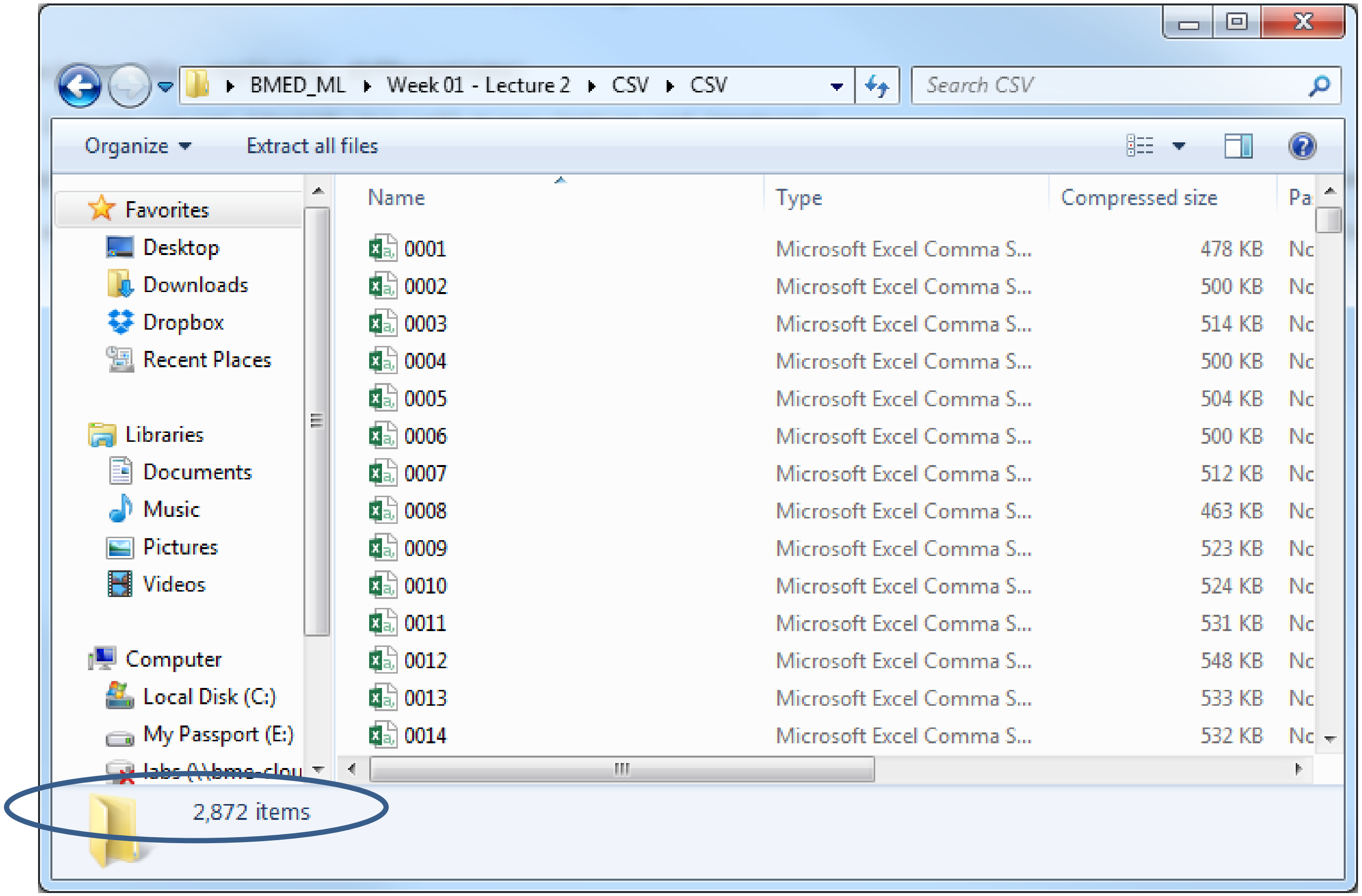
- For each tube, 7 channels are measured
  - FSC, SSC
  - 5 protein markers

- Normal/AML class labels of 179 samples are given

- Task: predict the class labels of the remaining 180 samples

Data

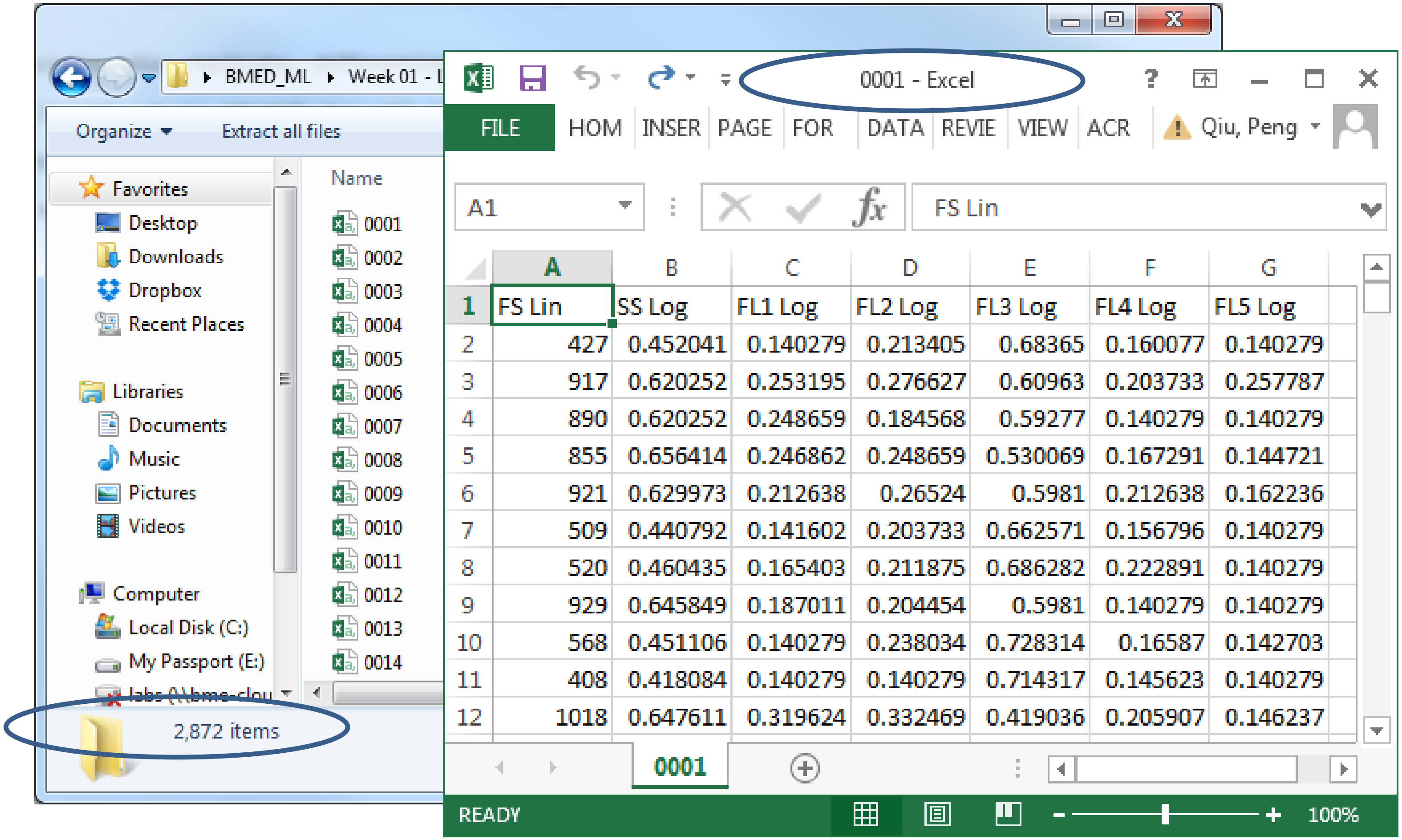
- 2872 data files in total (359 subjects * 8 tubes)
  http://pengqiu.gatech.edu/MLB/CSV.zip



Name	Type	Compressed size	Pa
0001	Microsoft Excel Comma S...	478 KB	Nc
0002	Microsoft Excel Comma S...	500 KB	Nc
0003	Microsoft Excel Comma S...	514 KB	Nc
0004	Microsoft Excel Comma S...	500 KB	Nc
0005	Microsoft Excel Comma S...	504 KB	Nc
0006	Microsoft Excel Comma S...	500 KB	Nc
0007	Microsoft Excel Comma S...	512 KB	Nc
0008	Microsoft Excel Comma S...	463 KB	Nc
0009	Microsoft Excel Comma S...	523 KB	Nc
0010	Microsoft Excel Comma S...	524 KB	Nc
0011	Microsoft Excel Comma S...	531 KB	Nc
0012	Microsoft Excel Comma S...	548 KB	Nc
0013	Microsoft Excel Comma S...	533 KB	Nc
0014	Microsoft Excel Comma S...	532 KB	Nc

Data

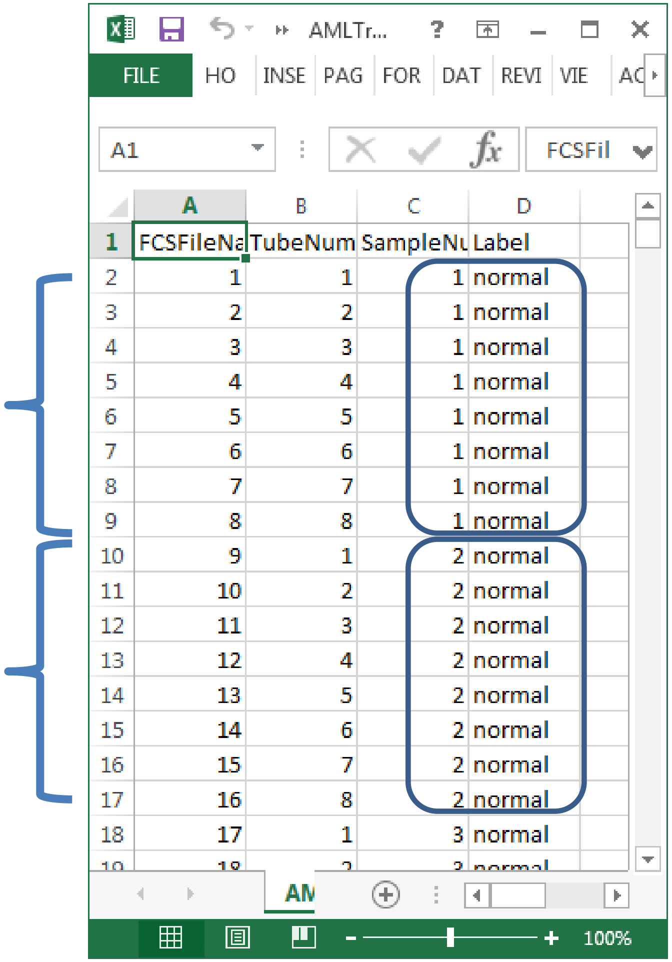
- 2872 data files in total (359 subjects * 8 tubes)
  http://pengqiu.gatech.edu/MLB/CSV.zip



	A	B	C	D	E	F	G
1	FS Lin	SS Log	FL1 Log	FL2 Log	FL3 Log	FL4 Log	FL5 Log
2	427	0.452041	0.140279	0.213405	0.68365	0.160077	0.140279
3	917	0.620252	0.253195	0.276627	0.60963	0.203733	0.257787
4	890	0.620252	0.248659	0.184568	0.59277	0.140279	0.140279
5	855	0.656414	0.246862	0.248659	0.530069	0.167291	0.144721
6	921	0.629973	0.212638	0.26524	0.5981	0.212638	0.162236
7	509	0.440792	0.141602	0.203733	0.662571	0.156796	0.140279
8	520	0.460435	0.165403	0.211875	0.686282	0.222891	0.140279
9	929	0.645849	0.187011	0.204454	0.5981	0.140279	0.140279
10	568	0.451106	0.140279	0.238034	0.728314	0.16587	0.142703
11	408	0.418084	0.140279	0.140279	0.714317	0.145623	0.140279
12	1018	0.647611	0.319624	0.332469	0.419036	0.205907	0.146237

Data

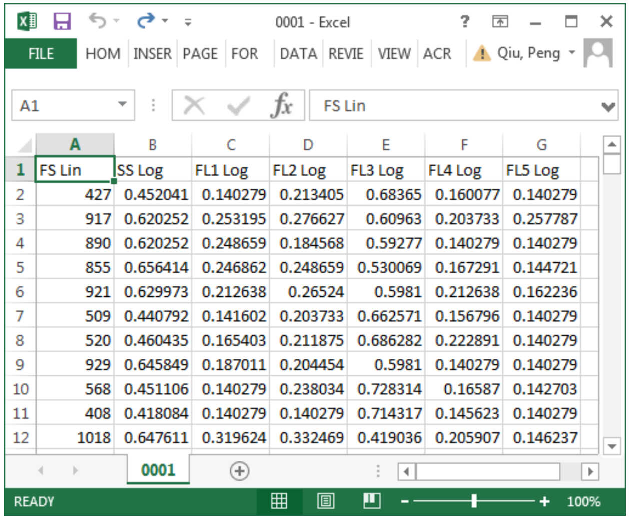
- Normal/AML class labels of 179 samples are given
  http://pengqiu.gatech.edu/MLB/AMLTraining.csv.zip



	A	B	C	D
1	FCSFileNa	TubeNum	SampleNu	Label
2	1	1	1	normal
3	2	2	1	normal
4	3	3	1	normal
5	4	4	1	normal
6	5	5	1	normal
7	6	6	1	normal
8	7	7	1	normal
9	8	8	1	normal
10	9	1	2	normal
11	10	2	2	normal
12	11	3	2	normal
13	12	4	2	normal
14	13	5	2	normal
15	14	6	2	normal
16	15	7	2	normal
17	16	8	2	normal
18	17	1	3	normal
19	18	2	3	normal

Data preprocessing: step 0

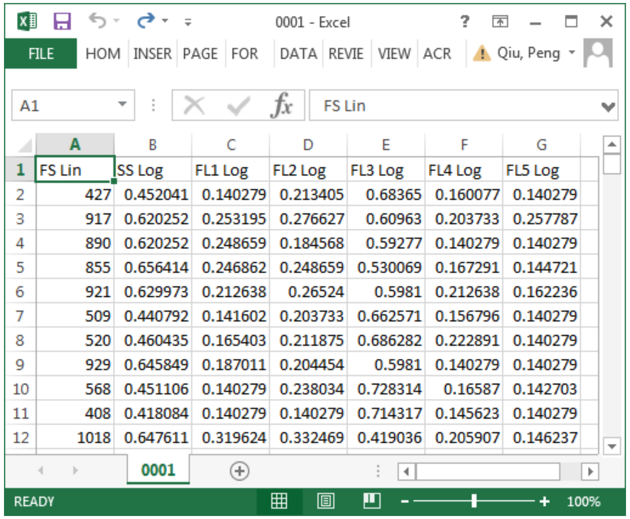
```
>> data = csvread('0001.CSV',1,0);
>> mean(data)

ans =

  663.9823    0.5544    0.2052    0.2023    0.5893    0.1818    0.1620

>> std(data)

ans =

  218.8498    0.0950    0.0522    0.0487    0.1073    0.0432    0.0261

>>
```



	A	B	C	D	E	F	G
1	FS Lin	SS Log	FL1 Log	FL2 Log	FL3 Log	FL4 Log	FL5 Log
2	427	0.452041	0.140279	0.213405	0.68365	0.160077	0.140279
3	917	0.620252	0.253195	0.276627	0.60963	0.203733	0.257787
4	890	0.620252	0.248659	0.184568	0.59277	0.140279	0.140279
5	855	0.656414	0.246862	0.248659	0.530069	0.167291	0.144721
6	921	0.629973	0.212638	0.26524	0.5981	0.212638	0.162236
7	509	0.440792	0.141602	0.203733	0.662571	0.156796	0.140279
8	520	0.460435	0.165403	0.211875	0.686282	0.222891	0.140279
9	929	0.645849	0.187011	0.204454	0.5981	0.140279	0.140279
10	568	0.451106	0.140279	0.238034	0.728314	0.16587	0.142703
11	408	0.418084	0.140279	0.140279	0.714317	0.145623	0.140279
12	1018	0.647611	0.319624	0.332469	0.419036	0.205907	0.146237

Data preprocessing: step 0

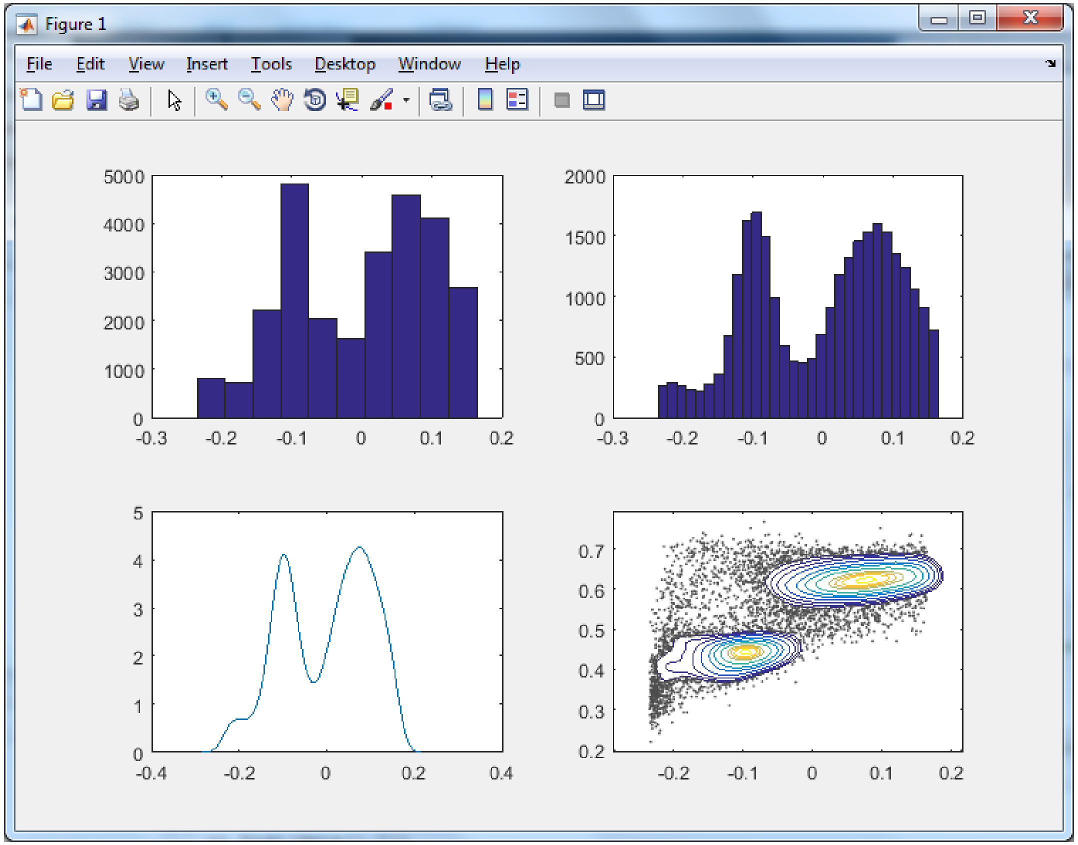
```
>> data(:,1) = data(:,1)-mean(data(:,1));
>> data(:,1) = data(:,1)/std(data(:,1))*0.1;
>> mean(data)

ans =

    0.0000    0.5544    0.2052    0.2023    0.5893    0.1818    0.1620

>> std(data)

ans =

    0.1000    0.0950    0.0522    0.0487    0.1073    0.0432    0.0261

>>
```



	A	B	C	D	E	F	G
1	FS Lin	SS Log	FL1 Log	FL2 Log	FL3 Log	FL4 Log	FL5 Log
2	427	0.452041	0.140279	0.213405	0.68365	0.160077	0.140279
3	917	0.620252	0.253195	0.276627	0.60963	0.203733	0.257787
4	890	0.620252	0.248659	0.184568	0.59277	0.140279	0.140279
5	855	0.656414	0.246862	0.248659	0.530069	0.167291	0.144721
6	921	0.629973	0.212638	0.26524	0.5981	0.212638	0.162236
7	509	0.440792	0.141602	0.203733	0.662571	0.156796	0.140279
8	520	0.460435	0.165403	0.211875	0.686282	0.222891	0.140279
9	929	0.645849	0.187011	0.204454	0.5981	0.140279	0.140279
10	568	0.451106	0.140279	0.238034	0.728314	0.16587	0.142703
11	408	0.418084	0.140279	0.140279	0.714317	0.145623	0.140279
12	1018	0.647611	0.319624	0.332469	0.419036	0.205907	0.146237

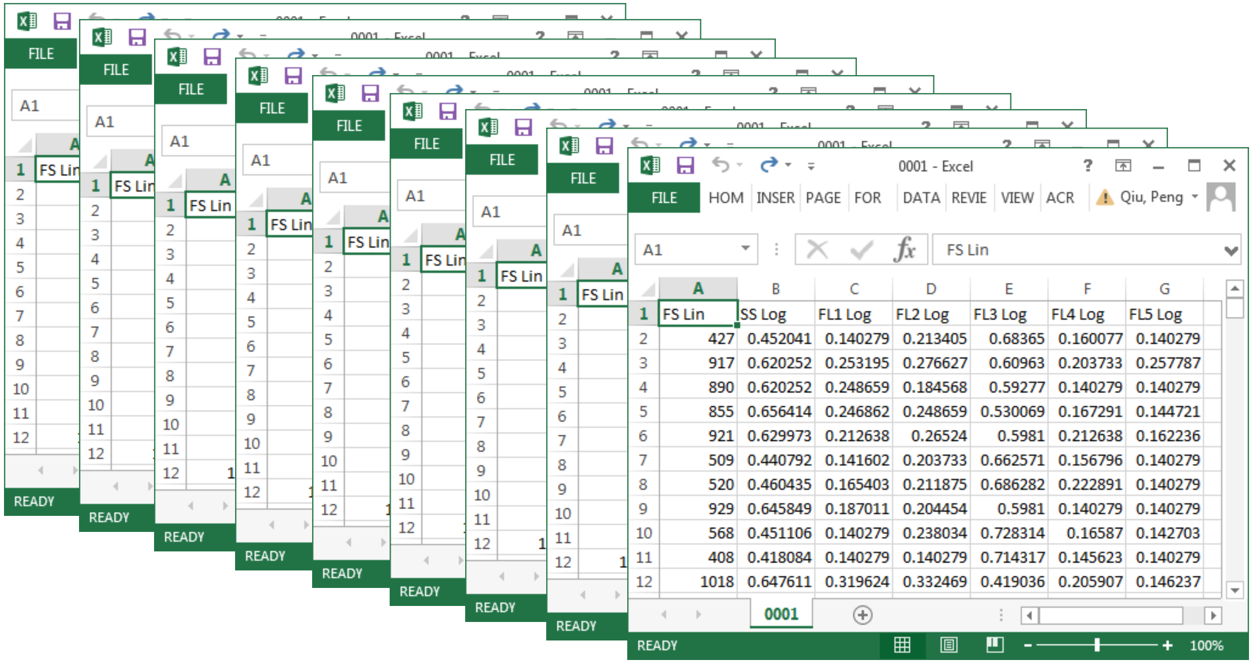
1D and 2D visualization

```
>> subplot(2,2,1); hist(data(:,1))
>> subplot(2,2,2); hist(data(:,1),30)
>> subplot(2,2,3); ksdensity(data(:,1))
>> subplot(2,2,4); FlowJo_contour2D(data(:,1),data(:,2),10)
```



More data preprocessing ???

Subject 1:

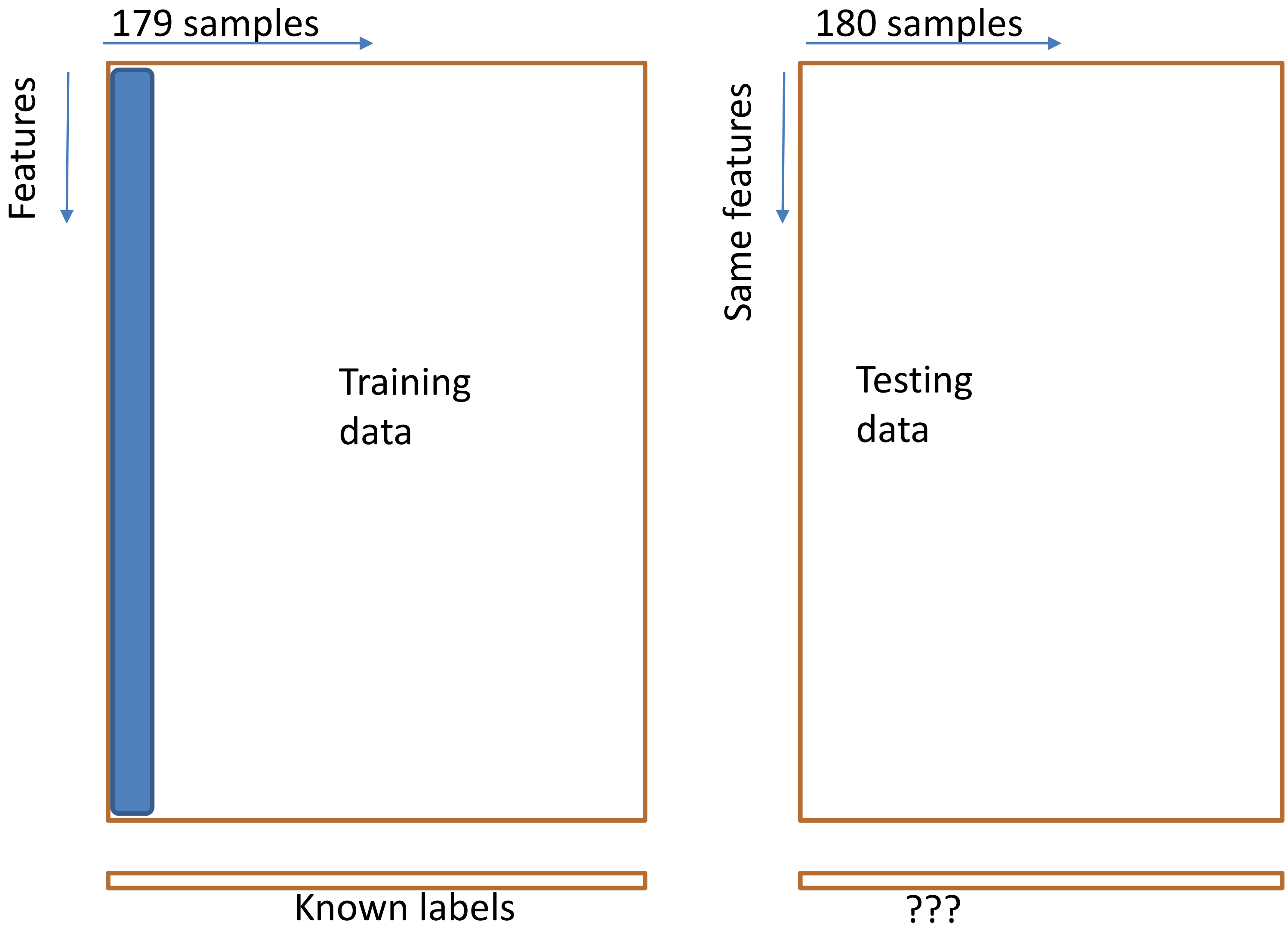


	A	B	C	D	E	F	G
1	FS Lin	SS Log	FL1 Log	FL2 Log	FL3 Log	FL4 Log	FL5 Log
2	427	0.452041	0.140279	0.213405	0.68365	0.160077	0.140279
3	917	0.620252	0.253195	0.276627	0.60963	0.203733	0.257787
4	890	0.620252	0.248659	0.184568	0.59277	0.140279	0.140279
5	855	0.656414	0.246862	0.248659	0.530069	0.167291	0.144721
6	921	0.629973	0.212638	0.26524	0.5981	0.212638	0.162236
7	509	0.440792	0.141602	0.203733	0.662571	0.156796	0.140279
8	520	0.460435	0.165403	0.211875	0.686282	0.222891	0.140279
9	929	0.645849	0.187011	0.204454	0.5981	0.140279	0.140279
10	568	0.451106	0.140279	0.238034	0.728314	0.16587	0.142703
11	408	0.418084	0.140279	0.140279	0.714317	0.145623	0.140279
12	1018	0.647611	0.319624	0.332469	0.419036	0.205907	0.146237

Subject 2: ⋮
⋮

⋮ ⋮
⋮ ⋮

Data in supervised setting



179 samples

Features

Training data

Known labels

180 samples

Same features

Testing data

???